Structure of a *Bacillus subtilis* endo- β -1,4-glucanase gene

Ron M. MacKay, Amy Lo, Gordon Willick, Michael Zuker, Stephen Baird, Michael Dove, Francesco Moranelli and Verner Seligy*

Molecular Genetics Section, Division of Biological Sciences, National Research Council of Canada, 100 Sussex Dr., Ottawa, Ontario K1A 0R6, Canada

Received 19 June 1986; Revised and Accepted 22 October 1986

ABSTRACT

The nucleotide sequence of the portion of a *Bacillus subtilis* (strain PAP115) 3 kb Pst I fragment which contains an endo- β -1,4-glucanase gene has been determined. This gene encodes a protein of 499 amino acid residues (M_r = 55,234) with a typical *B. subtilis* signal peptide. *Escherichia coli* which has been transformed with this gene produces an extracellular endoglucanase with an amino-terminus corresponding to the thirtieth encoded amino acid residue. The gene is preceded by a cryptic reading frame with a rho-independent terminator structure, and itself has such a structure in the immediate 3'-flanking region. We have also identified, in the 5'-flanking region, nucleotide sequences which resemble promoter elements recognized by *Bacillus* RNA polymerase $E\sigma^{32}$. Comparison of the encoded amino acid sequence to other known β -glucanases reveals a small region of similarity to the encoded protein of the *Clostridium thermocellum* celB gene. These similar regions may contain substrate-binding and/or catalytic sites.

INTRODUCTION

Bacillus species are well known for their abilities to produce one, or more, of a variety of extracellular enzymes capable of hydrolyzing polysaccharides, proteins, nucleic acids, lipids and antibiotics (1). Several *Bacillus subtilis* strains have been identified which produce endo- β -1,4-glucanases or cellulolytic endoglucanases, and the genomes of these have been used as sources for the molecular cloning of the respective genes (2-5). No similarity is evident between patterns of restriction enzyme cleavage of these cloned *B. subtilis* DNAs, indicating differences in gene structure between strains. Interest in our laboratory in the *B. subtilis* strain PAP115 endoglucanase derives, in part, from the potential utility of this activity in industrial applications (i.e., degradation of cellulosic wastes). We have previously reported on the identification of a DNA fragment from this strain which contains all the information needed for the expression of extracellular endoglucanase in transformed *Escherichia coli*, and with modification has been transiently expressed in two species of yeast, *Saccharomyces cerevisiae* and *Schwanniomyces alluvius* (5). The gene appears to be a single copy gene in *B.*

subtilis. Here we describe the structure of this gene and flanking regions, the encoded protein, and the amino-terminus of the protein produced by E. coli.

MATERIALS AND METHODS

Enzymes and reagents

Restriction enzymes, T⁴ DNA ligase, the exonuclease Bal-31 and the Klenow fragment of E. coli DNA polymerase were from Bethesda Research Laboratories or Pharmacia. [α -³²P]dATP (approx. 700 Ci/mmol) was purchased from Amersham or ICN. Deoxy- and dideoxynucleotides were obtained from Boehringer-Mannheim. Synthetic oligonucleotides used as sequencing primers were from Pharmacia.

Molecular cloning and DNA sequence determination

The isolation of the B. subtilis PAP115 endo- β -1,4-glucanase gene and flanking regions as a 3 kb Pst I insert in the vector pUC8 has been described (5). Recombinant plasmid pC6.5 or pC6.3 was prepared from transformed E. coli strain HB101 according to Maniatis *et al.* (6). The insert fragment was excised with Pst I, resolved on a 0.7% agarose gel, and eluted from the gel by electrophoresis (6). Methods for the subcloning of restriction enzyme digests of this fragment, and of Bal-31 digested pC6.5 and pC6.3 (5), into the vector M13mp9 have been described (6,7). Transformations of E. coli strain JM103 were done as per Hanahan (8). The procedure for testing the expression of JM103 transformants are described in Fig. 1 and elsewhere (5,9). Determination of nucleotide sequences was performed as described by Messing (7). Fig. 2 indicates the strategy used to determine the nucleotide sequence reported in Fig. 3. Except for 58 bp of the right-hand terminus, both the sequence reported and the complement have been determined with subclones of opposite polarity.

Amino-terminus sequencing of the extracellular endoglucanase

Culture supernatants of E. coli strain JM103 transformed with pC6.3 (5) were concentrated by ethanol precipitation (60%, v/v) and loaded onto a phenyl-sepharose (Pharmacia) column (1.5 x 46 cm), using a 2.5 to 0M gradient of sodium chloride, buffered at pH 5.0 with 50 mM sodium acetate. The active fractions (as determined by reducing sugar assay; see ref. 9) were dialyzed, concentrated by filtration with an Amicon YM-10 membrane, and further purified by application to a Vydac TP phenyl reverse phase column (4.6 x 150 mm, 5 μ m). The column was eluted with a linear gradient of 0-100% of 22% propanol/15% butanol in 0.1 M ammonium acetate, pH 5.2. Fractions containing

Figure 1. Expression of cloned *Bacillus subtilis* endo- β -1,4-glucanase gene. A. Congo Red plate assay (5) showing distribution of extracellular (a,d), periplasmic (b,e) and intracellular (c,f) cellulase activity expressed from the 2.8 kbp HindIII fragment of pC6.3 (see Fig. 2) cloned into M13mp9 in same (pSB0; a-c) and opposite (pSB1; d-f) orientation to Lac Z gene induced with IPTG. Corresponding activity measured by reducing sugar assay (9) are: a, 0.33; b, 0.05; c, 0.04; d, 0.20; e, 0.04; f, 0.02. B. Time course of CM-cellulase hydrolytic activity measured by a Cannon viscometer (100-R714) at 30°C. Reaction contained: 0.5 ml of extracellular culture from *E. coli* JM103 (--) or JM103 transformed with pC6.3 (—) mixed with 9.5 ml of 0.5% CM-cellulose (Hercules) in 0.1 M Na Acetate pH 5.0. Viscosity was measured with identical samples.

Figure 2. Restriction map and nucleotide sequencing strategy for the *Bacillus subtilis* 3.1 kb Pst I fragment containing an endo- β -1,4-glucanase gene. All restriction enzyme cleavage sites are those predicted by the nucleotide sequence. The assignment of restriction sites in the leftward 0.67 kb of the 3.1 kbp fragment is tentative, due to the preliminary nature of sequencing data in this region. The large arrow pointing rightwards indicates the location and polarity of the endoglucanase-encoding open reading frame. The smaller arrows beneath the restriction maps depict the strategy for the subcloning of this fragment into the vector M13mp9 and subsequent nucleotide sequencing. Arrows without dashed lines connecting to restriction maps are subclones generated by exonuclease Bal-31 deletion of the parental clones (see Materials and Methods).

cellulase activity were pooled and examined on a 12.5% SDS-polyacrylamide gel (10). Sequencing was done with a Beckman System 890 sequencer.

Analysis of nucleotide sequence data

The encoded amino acid sequence of the B. subtilis endoglucanase was compared to other published endoglucanase sequences with the MATCH program developed by M. Zuker of the Biomathematics Section of the NRC of Canada, utilizing the Needleman and Wunsch algorithm (11,12) with identities scoring 1 point, non-identities scoring 0 points and gap penalties of $2 + 0.1(\text{gap size})$. Free energies for secondary structure configurations of nucleic acid sequences were calculated according to rules summarized by Salser (13).

RESULTS

Fig. 1A demonstrates the cellulolytic expression of E. coli JM103 transformed with the M13mp9 vector containing the endoglucanase gene in either the same (pSB0) or in the opposite (pSB1) orientation to the lac Z gene. The results of the viscometry assays provided in Fig. 1B further establish the endocellulolytic nature of the B. subtilis enzyme. These results are consistent with previous work (5) which showed that this fragment contains a gene with a functional promoter capable of transforming E. coli into a producer of extracellular endoglucanase activity. This same work also demonstrated that the polarity of transcription from this promoter is from the left to the right in the nucleotide sequence in Fig. 3. The downward expression of cellulase activity, determined by reducing sugar assay (see legend of Fig. 1 for data), when the lac Z and putative cellulase promoters are facing opposite each other is further consistent with previous work. The largest open reading frame we were able to find in either polarity, in the entire 3 kb fragment, is presented in Fig. 3 as encoding the amino acid sequence of the endoglucanase. Amino acid sequence analysis of the amino-terminus of the extracellular enzyme (produced by transformed E. coli) confirmed this assignment, the first fifteen residues of this protein being those of codons 30-44 of the identified open reading frame.

DISCUSSION

Structure of the gene and flanking regions

The endoglucanase gene identified in the nucleotide sequence of the B. subtilis PAP115 3 kb Pst I fragment encodes a protein of 499 amino acid residues (Fig. 3). We have indicated the methionine codon beginning at nucleotide position 731 to be the translation initiator. The first in-frame nonsense codon occurring 5' to this methionine codon begins at nucleotide

GCATCCAAGT CACACAGGCT GCTCTTAAC^{3'}

ATCACATTGG AGGCTGGGCG CCGACTTTCA^{6'} TCGCAGTCGC AATGTTCTTG TTTCATTCA GTTCAGTTG^{10'}

CGGCAACTAT TATTATGGCG AGACAAACAT^{13'} TGAATTTATT AAAACAAGCA AAACATGGCT GAACATTAC^{17'}

CGTATCGCTG TTATTGCTAT GGTGTGTAT^{20'} GGATCTTTAT CAGGCTTCCA AATCGTTTGG GATATGGCG^{24'}

ACCTCTTTAT GGGTATCATG GCGCTGATCA^{27'} ACTTAATTGT GATTGCGCTG CTGTCAAACG TTGCTTACA^{31'}

AGTGTATAAA GATTACGCGA AACAGCGTAA^{34'} GCAAGGACTT GATCCTGTGT TTAAGCGGAA AAACATCCCA^{38'}

GGGCTGAAAA ACGCTGAAAC ATGGGAAGAT^{41'} GAGAAACAAG AAGCATAATA TAAATATAAA ACAAAGCTGC^{45'}

ATTCAATAGT TGAATGCAGC TTTTTCATTA^{48'} TTGGAATAAA ATTTAATTTT TCGACTCGAA TGAATAATCA^{52'}

CGTGTAAGT CCAATTTCAG TCCAGTTTTC^{55'} TTTGTTCTAT ATGTGTCATG TGTGTCTTAT TCATTGGAG^{59'}

TAGAGCAAAT TGACAGGCTT TTAACCTCC^{62'} CCAAAACAAG AAATTAGGTT GATAGACAAT CATGAGAAA^{66'}

ATTTTACAA TGAGTTCGTG CTCATAAGAA^{69'} GTGTAGAGCC AAAATGATGC GAAGGAGGAA AAGATCAGAT^{73'}

ATG AAA CGG TCA ATC TCT ATT TTT ATT ^{76'} TGT TTA TTG ATT ACG TTA TTG ACA ATG ^{80'} GGC
Met Lys Arg Ser Ile Ser Ile Phe Ile Thr Cys Leu Leu Ile Thr Leu Leu Thr Met Gly
5 15

GGC ATG ATA GCT TCG CCG GCA TCA GCA ^{83'} GGG ACA AAA ACG CCA GTA GCC AAG AAT ^{87'} GGC
Gly Met Ile Ala Ser Pro Ala Ser Ala Ala Gly Thr Lys Thr Pro Val Ala Lys Asn Gly
25 35

CAG CTT AGC ATA AAA GGT ACA CAG CTC GTT ^{90'} AAC CGA GAC GGT AAA GCG GTA CAG CTG ^{94'} AAG
Gln Leu Ser Ile Lys Gly Thr Gln Leu Val Asn Arg Asp Gly Lys Ala Val Gln Leu Lys
45 55

GGG ATC AGT TCA CAC GGA TTG CAA TGG TAT ^{97'} GGA GAA TAT GTC AAT AAA GAC AGC TTA ^{101'} AAA
Gly Ile Ser Ser His Gly Leu Gln Trp Tyr Gly Glu Tyr Val Asn Lys Asp Ser Leu Lys
65 75

TGG CTG AGA GAT GAT TGG GGT ATC ACC ^{104'} GTT TTC CGT GCA GCG ATG TAT ACG GCA GAT ^{108'} GGC
Trp Leu Arg Asp Asp Trp Gly Ile Thr Val Phe Arg Ala Ala Met Tyr Thr Ala Asp Gly
85 95

GGT TAT ATT GAC AAC CCG TCC GTG AAA ^{111'} AAT AAA GTA AAA GAA GCG GTT GAA GCG GCA ^{115'} AAA
Gly Tyr Ile Asp Asn Pro Ser Val Lys Asn Lys Val Lys Glu Ala Val Glu Ala Ala Lys
105 115

GAG CTT GGG ATA TAT GTC ATC ATT GAC ^{118'} TGG CAT ATC TTA AAT GAC GGT AAT CCA AAC ^{122'} CAA
Glu Leu Gly Ile Tyr Val Ile Ile Asp Trp His Ile Leu Asn Asp Gly Asn Pro Asn Gln
125 135

AAT AAA GAG AAG GCA AAA GAA TTC TTC ^{125'} AAG GAA ATG TCA AGC CTT TAC GGA AAC ACG ^{129'} CCA
Asn Lys Glu Lys Ala Lys Glu Phe Phe Lys Glu Met Ser Ser Leu Tyr Gly Asn Thr Pro
145 155

AAC GTC ATT TAT GAA ATT GCA AAC GAA ^{132'} CCA AAC GGT GAT GTG AAC TGG AAG CGT GAT ^{136'} ATT
Asn Val Ile Tyr Glu Ile Ala Asn Glu Pro Asn Gly Asp Val Asn Trp Lys Arg Asp Ile
165 175

AAA CCA TAT GCG GAA GAA GTG ATT TCA ^{139'} GTT ATC CGC AAA AAT GAT CCA GAC AAC ATC ^{143'} ATC
Lys Pro Tyr Ala Glu Glu Val Ile Ser Val Ile Arg Lys Asn Asp Pro Asp Asn Ile Ile
185 195

ATT GTC GGA ACC GGT ACA TGG AGC CAG ^{146'} GAT GTG AAT GAT GCT GCC GAT GAC CAG CTA ^{150'} AAA
Ile Val Gly Thr Gly Thr Trp Ser Gln Asp Val Asn Asp Ala Ala Asp Asp Gln Leu Lys
205 215

```

GAT GCA AAC GTT ATG TAC GCA CTT CAT122 TAT GCC GGC ACA CAC GGC CAA TTT TTA182 CGG
Asp Ala Asn Val Met Tyr Ala Leu His Phe Tyr Ala Gly Thr His Gly Gln Phe Leu Arg
225 235

GAT AAA GCA AAC TAT GCA CTC AGC AAA188 GGA GCA CCT ATT TTT GTG ACA GAG TGG GGA190
Asp Lys Ala Asn Tyr Ala Leu Ser Lys Gly Ala Pro Ile Phe Val Thr Glu Trp Gly Thr
245 255

AGC GAC GCG TCT GGC AAT GGC GGT GTA188 TTC CTT GAT CAA TCG AGG GAA TGG CTG AAA187 TAT
Ser Asp Ala Ser Gly Asn Gly Gly Val Phe Leu Asp Gln Ser Arg Glu Trp Leu Lys Tyr
265 275

CTC GAC AGC AAG ACC ATT AGC TGG GTG188 AAC TGG AAT CTT TCT GAT AAG CAG GAA TCA188 TCC
Leu Asp Ser Lys Thr Ile Ser Trp Val Asn Trp Asn Leu Ser Asp Lys Gln Glu Ser Ser
285 295

TCA GCT TTA AAG CCG GGC GCA TCT AAA186 ACA GGC GGC TGG CGG TTG TCA GAT TTA TCT188 GCT
Ser Ala Leu Lys Pro Gly Ala Ser Lys Thr Gly Gly Trp Arg Leu Ser Asp Leu Ser Ala
305 315

TCA GGA ACA TTC GTT AGA GAA AAC ATT178 CTC GGC ACC AAA GAT TCG ACG AAG GAC ATT178 CCT
Ser Gly Thr Phe Val Arg Glu Asn Ile Leu Gly Thr Lys Asp Ser Thr Lys Asp Ile Pro
325 335

GAA ACG CCA TCA AAA GAT AAA CCC ACA178 CAG GAA AAT GGT ATT TCT GTA CAG TAC AGA181 GCA
Glu Thr Pro Ser Lys Asp Lys Pro Thr Gln Glu Asn Gly Ile Ser Val Gln Tyr Arg Ala
345 355

GGG GAT GGG AGT ATG AAC AGC AAC CAA188 ATC CGT CCG CAG CTT CAA ATA AAA AAT AAC188 GGC
Gly Asp Gly Ser Met Asn Ser Asn Gln Ile Arg Pro Gln Leu Gln Ile Lys Asn Asn Gly
365 375

AAT ACC ACG GTT GAT TTA AAA GAT GTC188 ACT GCC CGT TAC TGG TAT AAA GCG AAA AAC188 AAA
Asn Thr Thr Val Asp Leu Lys Asp Val Thr Ala Arg Tyr Trp Tyr Lys Ala Lys Asn Lys
385 395

GGC CAA AAC TTT GAC TGT GAC TAC GCG188 CAG ATT GGA TGC GGC AAT GTG ACA CAC AAG188 TTT
Gly Gln Asn Phe Asp Cys Asp Tyr Ala Gln Ile Gly Cys Gly Asn Val Thr His Lys Phe
405 415

GTG ACG TTG CAT AAA CCA AAG CAA GGT220 GCA GAT ACC TAT CTG GAA CTT GGA TTT AAA200 AAC
Val Thr Leu His Lys Pro Lys Gln Gly Ala Asp Thr Tyr Leu Glu Leu Gly Phe Lys Asn
425 435

GGA ACG TTG GCA CCG GGA GCA AGC ACA200 GGG AAT ATT CAG CTC CGT CTT CAC AAT GAT210 GAC
Gly Thr Leu Ala Pro Gly Ala Ser Thr Gly Asn Ile Gln Leu Arg Leu His Asn Asp Asp
445 455

TGG AGC AAT TAT GCA CAA AGC GGC GAT210 TAT TCC TTT TTC AAA TCA AAT ACG TTT AAA210 ACA
Trp Ser Asn Tyr Ala Gln Ser Gly Asp Tyr Ser Phe Phe Lys Ser Asn Thr Phe Lys Thr
465 475

ACG AAA AAA ATC ACA TTA TAT GAT CAA220 GGA AAA CTG ATT TGG GGA ACA GAA CCA AAT220 TAG
Thr Lys Lys Ile Thr Leu Tyr Asp Gln Gly Lys Leu Ile Trp Gly Thr Glu Pro Asn ***
485 495

TTAAGCTTTT TTTTGGCGGA CATCAGTAA220 CATGTCGGT TTTATCATCT AATAGGAGGT TTGCGTTGTG210
GATGAAATAC TGAAACAGTA TATGGTCTG220 TATAAAAAAA TGAGTAATAT GATAAATGGT CCCGACTATC210
CAGGTAAGGA AAAAGACATC CAGCATCA220 AAGATCAGAT CGAAGTTTAC GAAAACAGC TGCAG220

```

Figure 3. Nucleotide sequence of the sense strand of the *Bacillus subtilis* endo- β -1,4-glucanase gene and flanking regions. Small numerals above and larger numerals below each line denote nucleotide and amino acid positions of the encoded protein, respectively. The sequence reported extends from 0.67 kb to the rightward Pst I site in Fig. 2. Short arrows underline putative translation termination codons for cryptic open reading frames in the regions which flank the endoglucanase gene. Long arrows underline inverted repeats with the potential to form rho-independent transcription termination

structures. Brackets at nucleotide positions 643-648 and 666-671 indicate possible "-35" and "-10" promoter elements for the Bacillus RNA polymerase $E\sigma^{37}$. The primary structure of the amino-terminus of the extracellular endoglucanase produced by transformed Escherichia coli is denoted by the underlining of amino acid positions 30-44. The small arrow at nucleotide position 1738 indicates the point to which Bal-31 digestion has removed the carboxy-terminal third of this gene in a recombinant molecule which retains the ability to direct the synthesis of active extracellular endoglucanase (see Discussion).

position 671. Two other potential, in-frame methionine codons occur (tandemly, at nucleotide positions 704-709) between this nonsense codon and the indicated putative initiator codon. The lack of a nucleotide sequence appropriate for hydrogen-bonding to the 3'-terminus of B. subtilis 16S rRNA (14,15) (i.e., a ribosome-binding site) in the region immediately 5' to those potential tandem methionine codons, and the presence of such a sequence 5' to the methionine codon at nucleotide position 731, suggests that nucleotide positions 704-709 are not part of the translated portion of the mRNA. We also note that the distance (14 bp between the A residues at positions 716 and 731) between the putative 16S rRNA-binding site and the indicated initiator codon is within the bounds observed for a variety of other Bacillus genes (15).

Investigation of the region 5' to the gene revealed the presence of sequences which could function as "-35" and "-10" promoter elements for transcription by Bacillus RNA polymerase $E\sigma^{37}$ (15-17), the major holo-enzyme involved in transcription of most genes during logarithmic growth. These elements are the 5'-TAGACA-3' and 5'-TACAAT-3' beginning at positions 643 and 666, respectively, in Fig. 3. These elements are separated by 17 bp, a typical arrangement for $E\sigma^{37}$ promoters (15, 18-22). Immediately 5' to the suggested -35 element is an A- and T-rich region. Such a region is often found in this situation in other Bacillus promoters (15,18,22,23).

Wong et al. (24) have suggested that genes which encode carbohydrate-degrading enzymes may be transcribed specifically by the $E\sigma^{37}$ RNA polymerase, a Bacillus holo-enzyme which transcribes genes expressed only during the stationary phase of culture. While Robson and Chambliss (4) have found that the extracellular endoglucanase produced by B. subtilis DLG did not appear until the onset of stationary phase, production of this activity parallels cell growth in B. subtilis IF03034 (3) and B. subtilis PAP115 (5). The $E\sigma^{37}$ holo-enzyme recognizes -35 and -10 promoter elements of 5'-AGGNTTT-3' and 5'-GGAATTGTTT-3', respectively (16,25). We are unable to find nucleotide sequences substantially similar to these in the region flanking the

endoglucanase gene in Fig. 3, nor any sequences similar to the proposed -35 and -10 elements recognized as promoters by the other known minor Bacillus holo-enzymes, $E\sigma^{32}$, $E\sigma^{29}$ and $E\sigma^{28}$ (16,26,27). We conclude that this gene is probably not developmentally regulated on the manner suggested by Wong et al. (24).

We note that the 5'-flanking region of the endoglucanase gene contains an open reading frame which terminates at a TAA codon located at nucleotide positions 426-428 in Fig. 3, and which extends in the 5'-direction for as far as we have sequenced the 3 kb Pst I fragment. Just 3' to the termination codon, a perfect 14 bp inverted repeat occurs (with the potential to form a hairpin loop with $\Delta G = -20$ kcal/mole) followed by a short T-rich segment, strongly suggesting rho-independent termination of transcription of this gene (17,22,28-31). This observation predicts that a Bacillus transcription promoter should be present in the region (approx. nucleotide positions 480-710) between this terminator and the ribosome-binding site of the endoglucanase gene (see above).

Immediately 3' to the termination codon of the endoglucanase gene we find another rho-independent transcription termination structure, consisting of an inverted repeat of 8 bp (forming a hairpin with $\Delta G = -13.7$ kcal/mole) and flanked by T-rich segments. A cryptic open reading frame, extending from the right-ward end of the sequence in Fig. 3, terminates in two tandem nonsense codons (nucleotide positions 2277-2272) just to the right of the inverted repeat. This observation suggests that the terminator structure may be utilized by two genes of opposite polarity.

Structure of the encoded protein

The large open reading frame identified in Fig. 3 encodes a protein of 499 amino acid residues (calculated $M_r = 55,234$). This protein is similar in size to the intracellular endoglucanase produced by E. coli transformed with a B. subtilis strain DLG endoglucanase gene ($M_r = 51 \pm 0.9 \times 10^3$; see ref. 4).

Selection of the indicated codon as the initiator is supported by comparison of the putative signal peptide (necessary for extracellular transport; see ref. 32-34) to other known B. subtilis signal peptides (see Fig. 4). A signal peptide beginning at amino acid position 1 in Fig. 3 would contain two basic residues (Lys2 and Arg3) in the hydrophilic amino-terminus (positions 1-6). The remainder of the peptide has an 18-residue hydrophobic core region which is rich in leucine and isoleucine residues, and a short region just before the signal peptide cleavage site which contains several

Endo- β -1,4-xylanase	(44)	MF ⁺⁺ K ⁺⁺ KN ⁺⁺ FLVGLSAA ⁺⁺ LSISLESATASA
α -Amylase	(45)	MFAKRFKTS ⁺⁺ LLPLFAGFLL ⁺⁺ LFYLVLAGPAA ⁺⁺ ASA
Levansucrase	(46)	MNIKKFAKQAT ⁺⁺ VLTF ⁺⁺ TALLAGGA ⁺⁺ TQAFA
Alkaline protease	(24)	MRSK ⁺⁺ KLWISLLFALT ⁺⁺ LIFTMAFSNMSAQA
Endo- β -1,3-1,4-glucanase	(37)	MPYLK ⁺⁺ RVLL ⁺⁺ LLVTGLFMSLFAVTATASA
Endo- β -1,4-glucanase		MKRSISIFITCL ⁺⁺ LITLLTMGGMIAS ⁺⁺ PASA

Figure 4. Comparison of signal peptides of *Bacillus subtilis* extracellular proteins. Sites of signal peptidase cleavage (i.e., the carboxy-terminii) are either those determined by amino-terminus sequencing of extracellular proteins (see references in parentheses) or those predicted by cleavage site recognition rules (33). The symbol + denotes basic residues in the hydrophilic leader regions. The brackets underline the hydrophobic core of each peptide.

alanine residues in a hydrophilic, but uncharged, environment.

The sequence of the amino-terminus of the extracellular endoglucanase produced by transformed *E. coli* (see Fig. 3) indicates that cleavage by the signal peptidase occurs precisely at the site predicted by previous work (33). Alanine, serine or glycine are required at the first and third positions to the amino-side of the cleavage site (Ala29 and Ala27 in Fig. 3), and there is usually a nearby proline residue (Pro26). The endoglucanase protein, lacking only its signal peptide, would consist of 470 amino acid residues with $M_r = 52,211$. Our previous observations (5) on the estimated molecular weights of active extracellular endoglucanases produced by transformed *E. coli* (M_r range of $30-49 \times 10^3$) and *B. subtilis* PAP115 (M_r range of $27-40 \times 10^3$) suggests that in both species considerable processing or degradation of the primary translation product occurs, beyond signal peptide cleavage. Amino-terminus sequencing indicates that, in *E. coli* at least, these proteolytic events remove substantial portions of the carboxy-half of the protein.

Similarity to other endoglucanases

Nucleotide sequences have recently become available for two of the endo- β -1,4-glucanase genes of another gram-positive bacterium, *Clostridium thermocellum* (35,36) and the endo- β -1,3-1,4-glucanase of *B. subtilis* (37). We have compared the *B. subtilis* PAP115 protein to the proteins encoded by each of these genes, and (as negative controls) to two non-homologous proteins of similar size (hog pancreatic and *Aspergillus oryzae* α -amylases; see citations in ref. 38). Comparison of the *B. subtilis* endo- β -1,4-glucanase to both the *B. subtilis* endo- β -1,3-1,4-glucanase and the *C.*

```

C. thermocellum  ** LVADKGINVVRMPIATDLLYAWSQGIYPPSTDTSYNNPALAGLSYELFNFMLENFKRVGIKVILDVHSPETDNQGHN 165
                  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
B. subtilis      1 RDDWGITVFRAAMYT---ADGGVIDNPS-----VKNKVKEAVEAAKELGIYVIDWHILNDGNPNQN 141
                  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

```

Figure 5. Alignment at similar regions of primary structure of the endo- β -1,4-glucanases of *Clostridium thermocellum* (celB gene) and *Bacillus subtilis*. Asterisks emphasize identical positions in this alignment. Dashed lines indicate where it is necessary to introduce gaps in the *B. subtilis* sequence. Note that each of these regions is approximately the same distance (in terms of primary structure) from the amino terminus of their respective unprocessed proteins.

thermocellum endo- β -1,4-glucanase encoded by the celA gene revealed no significant similarities between these proteins. Comparison of the endo- β -1,4-glucanases of *B. subtilis* and the celB gene of *C. thermocellum* revealed no significant similarity overall; however, we were able to identify a short region of relatively high similarity. The alignment of amino acids 82-141 of the *B. subtilis* protein to amino acids 88-165 of the *C. thermocellum* celB protein is shown in Fig. 5. Comparison of either protein to the α -amylases, yielded no regions of similarity as extensive as those in Fig. 5, particularly the fifteen positions corresponding to *B. subtilis* amino acids 117 to 131. Comparison of bacterial and eukaryotic α -amylases (38-41) have shown that active site residues are found in small, highly-conserved regions. The regions identified in Fig. 5 may contain some of the active site residues of the *B. subtilis* endoglucanase. The location of functionally important regions near the amino-terminus is consistent with our observation (data not shown) that a mutant gene which has suffered deletion of that portion of the gene corresponding to the carboxy-third of the protein (see Fig. 3) still directs production of active extracellular endoglucanase in *E. coli*.

We have been unable to find any similarity between the suggested (42) active site of the endo- β -1,4-glucanase of the basidiomyceteous fungus *Schizophyllum commune* and the regions of the bacterial endo-glucanases depicted in Fig. 5. If these regions do contain active site residues, it would appear that the gram-positive eubacteria and the eukaryotes differ greatly in the structure of these sites.

The lack of overall similarity between the *C. thermocellum* celA and celB endoglucanases and that of *B. subtilis* is somewhat disturbing in view of the observations of Pétré *et al.* (43) that both of the respective *Clostridium* genes (or DNA sequences adjacent to these genes) hybridize to genomic DNA of *B. subtilis* strain QB666, indicating that these genes may have homologues in

some B. subtilis strains. If the strain we have (PAP115) does contain homologues to celA and celB, the gene we have identified is certainly not one of these.

Elucidation of the evolutionary relationships between eubacterial endoglucanases and definition of functionally important regions of these proteins requires the accumulation of more protein and/or gene structure data, and studies of the activity of mutated genes.

ACKNOWLEDGEMENTS

The authors wish to thank R. Staden for providing the computer program used to align subclone nucleotide sequences, and C. Roy and M. Yaguchi for determination of the structure of the amino-terminus of the endoglucanase protein. This is NRCC publication no. 25953.

*To whom correspondence should be addressed

REFERENCES

1. Priest, F.G. (1977) Bacteriol. Rev. **41**, 711-753.
2. Sashihara, N., Kudo, T. and Horikoshi, K. (1984) J. Bacteriol. **158**, 503-506.
3. Koide, Y., Nakamura, A., Uozumi, T. and Beppu, T. (1986) Agric. Biol. Chem. **50**, 233-237.
4. Robson, L.M. and Chambliss, G.H. (1986) J. Bacteriol. **165**, 612-619.
5. Seligy, V., Baird, S., Dove, M., Lo, A., MacKay, R., Willick, G., Bernier, R. and Desrochers, M. (1987) In Stewart, G.G., Russell, I., Klein, R.D. and Hiebish, R.R. (eds), The Biochemistry and Molecular Biology of Industrial Yeasts, Uniscience Series, CRC Press, Boca Raton (in press).
6. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning, A Laboratory Manual, Cold Spring Harbour Laboratory, Cold Spring Harbour.
7. Messing, J. (1983) In Wu, R., Grossman, L. and Moldave, K. (eds), Methods in Enzymology, Academic Press, New York, Vol. 101, pp. 20-78.
8. Hanahan, D. (1983) J. Mol. Biol. **166**, 557-580.
9. Miller, G.L. (1959) Anal. Chem. **31**, 426-428.
10. Laemmli, U.K. (1970) Nature **227**, 680-685.
11. Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. **48**, 443-453.
12. Hogeweg, P. and Hesper, B. (1984) J. Mol. Evol. **20**, 175-186.
13. Salser, W. (1972) Cold Spring Harbour Symp. Quant. Biol. **42**, 985-1002.
14. Band, L. and Henner, D.J. (1984) DNA **3**, 17-21.
15. Moran, C.P., Jr., Lang, N., LeGrice, S.F.J., Lee, G., Stephens, M., Sonenshein, A.L., Pero, J. and Losick, R. (1982) Mol. Gen. Genet. **186**, 339-346.
16. Johnson, W.C., Moran, C.P., Jr. and Losick, R. (1983) Nature **302**, 800-804.
17. Gitt, M.A., Wang, L.-F. and Doi, R.H. (1985) J. Biol. Chem. **260**, 7178-7185.
18. Fujita, Y. and Fujita, T. (1986) Nucl. Acids Res. **14**, 1237-1252.
19. Shimotsu, H. and Henner, D.J. (1986) Proc. Natl. Acad. Sci. USA **81**, 6315-6319.
20. Dobinson, K.F. and Spiegelman, G.B. (1985) J. Biol. Chem. **260**, 5950-5955.

21. Wang, P.-Z. and Doi, R.H. (1984) J. Biol. Chem. **259**, 8619-8625.
22. Lehtovaara, P., Ulmanen, I. and Palva, I. (1984) Gene **30**, 11-16.
23. Banner, C.D.B., Moran, C.P., Jr. and Losick, R. (1983) J. Mol. Biol. **168**, 351-365.
24. Wong, S.-L., Price, C.W., Goldfarb, D.S. and Doi, R.H. (1984) Proc. Natl. Acad. Sci. USA **81**, 1184-1188.
25. Tatti, K.M. and Moran, C.P., Jr. (1984) J. Mol. Biol. **175**, 285-297.
26. Hay, R.E. and Moran, C.P., Jr. (1985) In Hoch, J.A. and Setlow, P. (eds), Molecular Biology of Microbial Differentiation, American Society for Microbiology, Washington, pp.170-175.
27. Tatti, K.M., Kenney, T.J., Hay, R.E. and Moran, C.P., Jr. (1985) Gene **36**, 151-157.
28. Holmes, W.M., Platt, T. and Rosenberg, M. (1983) Cell **32**, 1029-1032.
29. Vasantha, N., Thompson, L.D., Rhodes, C., Banner, C., Nagle, J. and Filpula, D. (1984) J. Bacteriol. **159**, 811-819.
30. Takkinen, K., Pettersson, R.F., Kalkkinen, N., Palva, I., Söderlund, H. and Kääriäinen, L. (1983) J. Biol. Chem. **258**, 1007-1013.
31. Mongkolsuk, S., Duvall, E.J. and Lovett, P.S. (1985) Gene **37**, 83-90.
32. Oliver, D. (1985) Ann. Rev. Microbiol. **39**, 615-648.
33. Dierstein, R. and Wickner, W. (1986) EMBO J. **5**, 427-431.
34. Mézes, P.S.F. and Lampen, J.O. (1985) In Dubnau, D.A. (ed), The Molecular Biology of the Bacilli, Academic Press, Orlando, Vol. II, pp.151-183.
35. Béguin, P., Cornet, P. and Aubert, J.-P. (1985) J. Bacteriol. **162**, 102-105.
36. Grépinet, O. and Béguin, P. (1986) Nucl. Acids Res. **14**, 1791-1799.
37. Murphy, N., McConnell, D.J. and Cantwell, B.A. (1984) Nucl. Acids Res. **12**, 5355-5367.
38. MacKay, R.M., Baird, S., Dove, M.J., Erratt, J.A., Gines, M., Moranelli, F., Nasim, A., Willick, G.E., Yaguchi, M. and Seligy, V.L. (1985) BioSystems **18**, 279-292.
39. Rogers, J.C. (1985) Biochem. Biophys. Res. Comm. **128**, 470-476.
40. Ihara, H., Sasaki, T., Tsuboi, A., Yamagata, H., Tsukagoshi, N. and Vdaka, S. (1985) J. Biochem. **98**, 95-103.
41. Nakajima, R., Imanaka, T. and Aiba, S. (1986) Appl. Microbiol. Biotechnol. **23**, 355-360.
42. Yaguchi, M., Roy, C., Rollin, C.F., Paice, M.G. and Jurasek, L. (1983) Biochem. Biophys. Res. Comm. **116**, 408-411.
43. Pétré, D., Béguin, P., Millet, J. and Aubert, J.-P. (1985) Ann. Inst. Pasteur/Microbiol. **136B**, 8-124.
44. Paice, M.G., Bourbonnais, R., Desrochers, M., Jurasek, L. and Yaguchi, M. (1986) Arch. Microbiol. **144**, 201-206.
45. Yamazaki, H., Ohmura, K., Nakayama, A., Takeichi, Y., Otozai, K., Yamasaki, M., Tamura, G. and Yamane, K. (1983) J. Bacteriol. **156**, 327-337.
46. Steinmetz, M., LeCoq, D., Aymerich, S., Gonzy-Tréboul, G. and Gay, P. (1985) Mol. Gen. Genet. **200**, 220-228.